

Deep Multi-Instance Multi-Label Learning for Image Annotation

Hai-Feng Guo^{*,†,¶}, Lixin Han^{*,†,¶}, Shoubao Su[†]
and Zhou-Bao Sun[§]

**College of Computer and Information
Hohai University, Nanjing 210024, P. R. China*

*†School of Computer Engineering, Jinling Institute of Technology
Nanjing 211169, P. R. China*

*‡State Key Laboratory of Novel Software Technology
Nanjing University, Nanjing 210093, P. R. China*

§Nanjing Audit University, Nanjing 211815, P. R. China
¶lhan@hhu.edu.cn

Received 29 March 2017

Accepted 27 June 2017

Published 18 August 2017

Multi-Instance Multi-Label learning (MIML) is a popular framework for supervised classification where an example is described by multiple instances and associated with multiple labels. Previous MIML approaches have focused on predicting labels for instances. The idea of tackling the problem is to identify its equivalence in the traditional supervised learning framework. Motivated by the recent advancement in deep learning, in this paper, we still consider the problem of predicting labels and attempt to model deep learning in MIML learning framework. The proposed approach enables us to train deep convolutional neural network with images from social networks where images are well labeled, even labeled with several labels or uncorrelated labels. Experiments on real-world datasets demonstrate the effectiveness of our proposed approach.

Keywords: Multi-instance multi-label learning; deep learning; convolutional neural network.

1. Introduction

Over the last few years, Multi-Instance Multi-Label learning (MIML)³⁴ has attracted a lot of attention. In contrast to traditional supervised learning which is concerned with dealing with problems where one instance is related only to a single label, in MIML, an example is represented by a bag of instances and associated with a set of labels. The goal in MIML is to learn a classifier that predicts the label set for an unseen bag of instances. Many real-world applications with multiple semantic

[¶]Corresponding author.

meanings can be formalized under an MIML learning framework. For example, an image is typically a bag; the segments in it are instances which are associated with a set of labels, such as sea, sunset or beach. Image annotation aims to learn the association between the visual feature and the labels. The goal is to develop approaches that can annotate a new image with some relevant labels. MIML image annotation is becoming more and more important since the number of images users upload to social networks is growing exponentially, many of them are easy to obtain and free to use. Meanwhile, most images have not been labeled or weakly labeled, which makes the images hard to search and index. In this paper, we are interested to tackle the MIML image annotation problems.

With the development of social network and digital photography, studies to understand image semantic started. The problem can be formulated as MIML learning framework. Numerous approaches for MIML image annotation have been proposed and applied to image.^{2,17,25,30-34} In recent years, deep convolutional neural networks (CNNs) have demonstrated a promising performance in image feature learning. CNN is a special type of neural network that utilizes specific network structures, such as convolutional layers, pooling layers or fully connected layers. The image annotation approach based on CNN can be regarded as two main components: one is the multiple-layer architecture composed of several layers that learns image representations from raw pixels; the other is the loss layer that propagates supervision cues back and fine-tunes the deep network to learn better representations for the specific tasks. Therefore, we attempt to incorporate MIML into deep learning framework and apply the learned visual knowledge to assist the task of image annotation.

The traditional approaches regard an image as one indiscrete entity and annotate the images with single label, which is not appropriate for practical applications, since the real-world images have more than one object or concept. In order to annotate images well, it is important that we handle MIML images with deep learning methods. In order to solve the disadvantages mentioned above, in our research, we focus on the MIML image annotation and propose an approach named MIMLCNNs which is based on deep learning method CNNs. As shown in Ref. 19, the larger and deeper the network is, the better the performance can be. However, with the growth of the network, the number of parameters increases significantly, which leads to requiring more training samples to prevent over-fitting. Actually, it is not feasible to obtain sufficient labeled images. At the same time, labeling a lot of images is problematic and time-consuming. This process is expensive and involves lots of ambiguous decisions.

In this paper, we attempt to learn the MIML problem in a deep learning manner. Briefly, the proposed approach employs convolutional neural network to handle the social network images. We consider that a certain label is associated with a part of some images and vice versa. The basic assumption is that the labels associated to the same image have the relatedness. Existing approaches^{17,34} are based on a simple

degeneration strategy and the other approaches^{30,32} tackle the problem directly in a regularization framework. The main purpose of this paper is to propose an effective approach to achieve higher accuracy prediction results and avoid a high computational cost of learning for social network datasets. In this paper, we make the following major contributions.

- (1) We incorporate deep learning into a supervised learning framework in a principled manner.
- (2) We propose an integrated framework to learn deep representations with MIML assumptions. The proposed approach can be conducted through experiments in image-level model and instance-level model.
- (3) We introduce the real-world datasets for MIMLCNNs learning. Experiments show that it achieves convincing performance.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed approach. Section 4 reports on the experiment results. Finally, we summarize and conclude the paper in Sec. 5.

2. Related Work

In the machine learning literature, during the past few years, a number of MIML learning approaches^{8,11,17,19,28–30,32,34} have been developed which in general can be divided into two categories. One way is to use the traditional approaches as the bridge. An initial attempt was made by Zhou *et al.*³⁴ They proposed MIML-SVM and MIML-BOOST to solve the problem by degenerating MIML to its equivalence in traditional supervised learning. Nam Nguyen proposed a powerful approach named SISL-MIML¹⁷ to deal with the MIML problem. Briefly, the algorithm seeks the best suitable single label belonging to the set of labels for all instances. Subsequently, the set of labels of a test example is determined by aggregating all labels of instances in the bag. The approaches are inspiring, however, neither did they consider the deep representations, nor did they consider the corresponding relations between images and labels. By the way, the approaches are very time-consuming. The other way is to formulate such problems as a joint one and model them in an integrated regularization framework. Zhou and Zhang presented M3MIML³² approach which considers the problem as a quadratic programming problem and implemented in its dual form. Later on, more and more approaches^{8,11,28,29} were developed. Some methods try to exploit the relations between instances and labels by relying on prior knowledge or counting the co-occurrence of labels in training data, whereas the situation is often unavailable and the over-fitting risk exists.

In recent years, the social networks allow users to upload images and describe the image content with tags. Most deep learning frameworks are in fully supervised settings. However, as discussed before, the images are highly weakly labeled or even unlabeled, weakly supervised learning is started to study using features learned with

deep representations. Specifically, Song *et al.*¹⁸ proposed a model based on CNN features for weakly supervised object localization, Xu *et al.*²⁷ proposed to use deep learning to compute features for multi-instance learning in medical imaging, Li *et al.*¹³ investigated this challenging problem by exploiting labeled and unlabeled data through a semi-parametric regularization and taking advantage of the multi-label constraints into the optimization. Later on, more and more efforts^{4-6,15,16,20,21,24,26} tried to solve the image annotation problem; these approaches all employed handcrafted features. More recently, in contrast to handcrafted features, the learnt features with CNN have been adopted to address multi-label problems. CNNs have outperformed existing hand-crafted features in many applications. Krizhevsky *et al.*¹⁰ proposed an approach in image classification task and conducted experiments on ILSVRC 2012 which consists of images from 1000 categories. Many efforts have focused on the designation or regularization methods of the structures of CNN,^{9,10,14} and achieved impressive performance on specific tasks. In terms of image annotation, Barnard *et al.*¹ presented some correspondence models on matching segmented images with associated text; Li *et al.*¹² proposed a framework for Internet images; Wang and Forsyth²³ made progress on jointly learning attributes and object classes via multi-instance learning.

In this paper, different from the above work, we propose an integrated framework to learn deep representations with MIML assumptions for the task of image annotation. One of our goals is to train the CNNs to extract features that represent the semantic similarities among images.

3. The MIMLCNNs Approach

In this section, we will review preliminaries to MIML learning which will be used throughout the paper (see Sec. 3.1). Then, we present our method for learning deep representations in a supervised manner, and give an introduction to the framework of MIMLCNNs (see Sec. 3.2).

3.1. Preliminaries

Different from traditional supervised learning in which training samples are given as pairs, in MIML learning framework, let $X = R^d$ denote a d -dimensional instance space and Y the set of class labels, a learning algorithm typically takes a set of labeled training examples $L = \{(x_1, y_1) \cdots (x_i, y_i) \cdots (x_n, y_n)\}$ as input, where $x_i = \{x_{i1}, \dots, x_{in_i}\}$ is an instances bag and n_i is the number of instances in x_i , $y_i = \{y_{i1}, \dots, y_{ik}\}$ is a k -dimensional label vector and $y_{ik} = [0, 1]$. For any image $x_i = \{x_{i1}, \dots, x_{in_i}\}$, $y_{ik} = 1$ indicates the membership associating x_i with the k th label. The goal of MIML is to learn a function of the form $f_{\text{MIML}} : 2^x \rightarrow 2^y$ which predicts a set of labels for an unseen example. Given that the MIML assumption lies generally in instances and labels, we therefore propose to exploit the relationships by incorporating MIML into a deep learning framework.

3.2. The framework

Considering the advances achieved by deep learning, it is a better choice to employ deep representations instead of a shallow model to solve MIML learning problem. Inspired by Ref. 26, we use deep CNN as the architecture for learning representation with MIML learning. The structure contains five convolutional layers, as shown in Fig. 1, followed by a pooling layer and three connected layers.

Here, P stands for a pooling layer, and C stands for a convolutional layer, and FC stands for a fully connected layer. Given a training sample, the network extracts layer-wise representations from the first convolutional layer to the output of the last fully connected layer, which can be regarded as high-level features of the input image. After obtaining the features, we can use the classification methods to classify the images into different categories.

In order to learn the MIML problem, we incorporate deep representation with MIML learning. A MIML CNN extracts representations of the instance and label bag:

$$h = \{h_{ij}\} \in R^{m \times n}, \quad i = 1, \dots, m; \quad j = 1, \dots, n,$$

in which each column is the representation of an instance, the aggregated representation of the bag for MIML is

$$\hat{h}_j = f \begin{pmatrix} h_{j11} & h_{j12} \cdots h_{j1m} \\ h_{j21} & h_{j22} \cdots h_{j2m} \\ h_{jn1} & h_{jn1} \cdots h_{jnm} \end{pmatrix}$$

where function f is the max layer, m is the number of instances and n is the number of labels. Followed by a soft-max layer, $FC8$ is transformed into a probability distribution for images of n categories, and here cross entropy is used to measure the prediction loss of the convolutional neural network. So, we have

$$p_i = \frac{\exp(\hat{h}_j)}{\sum_j \exp(\hat{h}_j)},$$

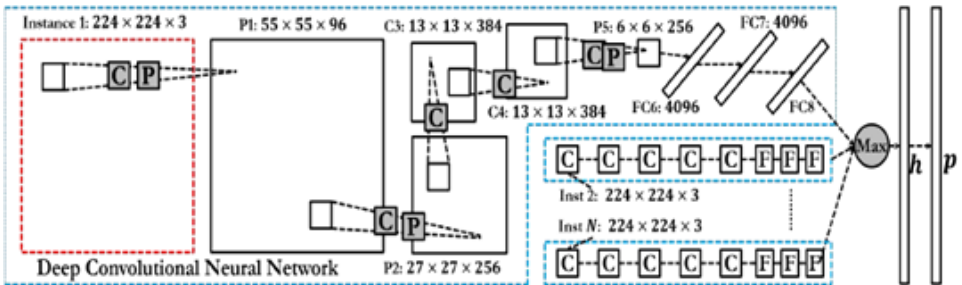


Fig. 1. The structure of the MIML CNNs framework.

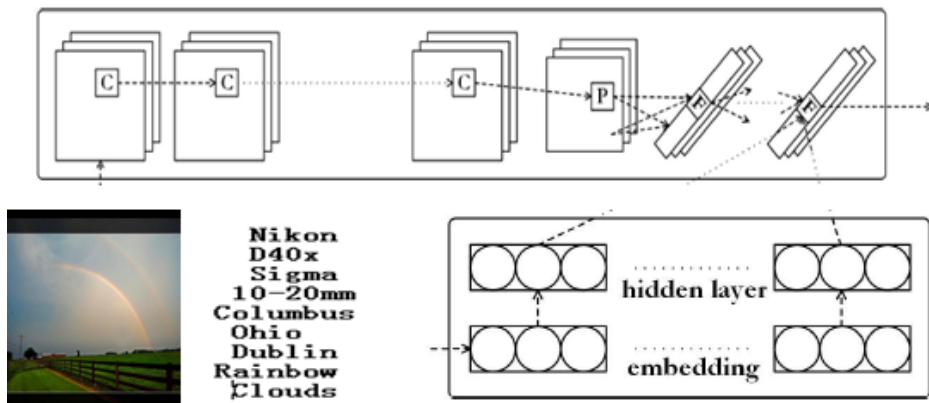


Fig. 2. Deep MIML learning framework.

$$L = - \sum_j t_j \log(p_j),$$

$$t = \left\{ t_j \mid t_j \in \{0, 1\}, j = 1, \dots, n, \sum_{j=1}^n t_j = 1 \right\}.$$

In order to minimize the loss function of deep MIML problem, we employ stochastic gradient descent for optimization, where the gradients of the connecting weights in each layer are calculated by a back propagation.

In this paper, we use a deep neural network formulation with MIML learning, the deep network contains one input layer, one hidden layer, and one output layer with soft-max. For instance-level learning, we build a joint MIML learning architecture to learn the instances and labels simultaneously. Specifically, we combine the outputs of image and labels understanding systems in the final fully connected layer, the framework is shown in Fig. 2.

4. Experiments

In this section, we conduct experiments of our deep learning framework on real-world datasets to validate the effectiveness of our approach. The proposed approach is implemented in MATLAB7.1. All the experiments are conducted on a Linux virtual machine with Intel processors (2.7 GHz) and 2 GB memory. We apply the proposed approach on image datasets for both image-level and instance-level annotations.

In this section, we conduct experiments of our deep learning framework on real-world datasets to validate the effectiveness of our approach. The proposed approach is implemented in MATLAB7.1. All the experiments are conducted on a Linux virtual machine with Intel processors (2.7 GHz) and 2 GB memory. We apply the

proposed approach on image datasets for both image-level and instance-level annotations.

4.1. Experiment setup

This section discusses the real-world datasets, experiment models, baseline approaches and metrics used in the experiments.

4.1.1. Datasets

The comparison is conducted on two real-world datasets, i.e., Microsoft Research Cambridge (MSRC v2)³ and MIRFLICKR-25000,⁷ all of the datasets can be available online. Tables 1 and 2 summarize the properties of each dataset used in the experiments.

The MIRFLICKR-25000 dataset, containing 25,000 images, 9,861 users, and more than 200,000 user-defined tags which were all retrieved from Flickr, is sufficiently large and tags assigned to images are the results of personal free tagging. Statistically, 455 users do not annotate any images, while 187 users annotate more than 10,000 favorite images, about 40% of the users annotate at least 500 favorite images, the average number of images annotated by the users of MIRFLICKR-25000 is 1263, and the average number of tags per image is 9. The statistics of the dataset are shown in Table 1.

Removing some concepts tags, such as geographical names, seasons and colors, the main tags are listed in Table 2.

The MIRFLICKR-25000 dataset is considered more challenging than the MSRC dataset as the objects are not centered and their appearances are more diverse. Moreover, users annotate the images arbitrarily, in other words many of the tags

Table 1. The statistics of dataset.

Dataset	Images	Tags/image	Tags	Users	MAX Tags	Medium	Users with no Tags
			(>= 20 images)			Tags	
MIRFLICKR-25000	25,000	8.94	1,386	9,861	1,263	301	2,128

Table 2. The main tags of the dataset.

Tag	Image Quantity	Tag	Image Quantity	Tag	Image Quantity	Tag	Image Quantity
Sky	845	Water	641	Portrait	623	Night	621
nature	596	Sunset	585	Clouds	558	Flower(s)	510/351
Beach	407	landscape	385	Street	383	Dog	372
architecture	354	Graffiti	335	Tree(s)	331/245	People	330
City	308	Sea	301	Sun	290	Girl	262
snow	256	Food	225	Bird	218	Sign	214
Car	12	Lake	199	building	188	River	175
naby	167	animal	164	streerart	184	urban	247

Table 3. The MSRC v2 dataset.

1	2	3	4	5	6	7	8	9	10
Building	Grass	Tree	Cow	House	sheep	Sky	Mountain	Aeroplane	Water
11	12	13	14	15	16	17	18	19	20
face	Car	Bicycle	Flower	Sign	bird	Book	Chair	Road	Cat
21	22	23							
dog	Body	Boat							

in MIRFLICKR-25000 dataset are irrelevant, and the images are weakly labeled. Although many of the labels are irrelevant to the image, some words actually provide more detailed and more informative descriptions than the category label does. This offers us an opportunity to obtain more specific image labels than user labeled tags.

The MSRC v2 dataset³ is a subset of the MSRC dataset named “v2” contains 591 images with 23 classes, and a total of 1,758 instances. There are around three labels per image on average. The label-cardinality is 5.0152 and the label-density is 0.2181. Each image is regarded as a bag and the label set is the union of the instance labels. Each instance is described by a 16-dimensional histogram of gradients and a 32-dimensional histogram of colors. The brief characteristic description of labels and corresponding class numbers are shown in Table 3.

The pair-wise label correlations between 23 labels in the MSRC dataset are illustrated in Fig. 3. The diagonal of the correlation matrix means the total number of each label, and the non-diagonal indicates the number of occurrences between each pair of labels.

4.1.2. Experiment models

In this paper, we conduct experiments on both datasets in image-level model and conduct experiments in instance-level model on the MSRC dataset. We consider

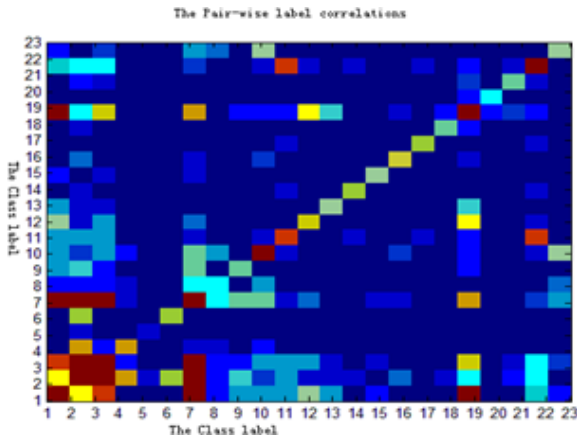


Fig. 3. The pair-wise label correlations.

conducting experiments to compare the performances of the proposed approach to existing approaches. The experiments are conducted to predict labels set for the unseen bags or instances.

4.1.3. Approaches

As reviewed in Sec. 2, there have been many approaches to solve the MIML problem. In this paper, we conduct experiments to compare the performance of the proposed approach MIMLCNNs with that of the other representative approaches: MIML-BOOST,³⁴ MIMLSVM,³⁴ SISL-MIML,¹⁷ M3MIML,³² MIMLKNN³⁰ and SIM.³ For a fair comparison, the optimal parameters for each approach are set according to the best settings as reported in Refs. 17, 30, 32 and 34. More precisely, as introduced in Ref. 34, the Gaussian kernel parameter of MIMLSVM is set to be 0.2 and the rounds of MIMLBOOST are set to be 25. For the M3MIML algorithm, the cost parameter C is set to the best values in the range $\{10^i | -4 \leq i \leq 4\}$ and γ is set to the default value of 1.0 as given in Ref. 32. For the MIMLKNN algorithm, the number of nearest neighbors is set to be 10 and the number of citers c is set to be 20. In addition, the SISL-MIML¹⁷ algorithm requires two parameters: the regularization constant $\lambda \in \{10^i | -4 \leq i \leq 4\}$ and the number of instances required to belong to a single label class $k \in \{1, 2, 3, 4\}$. For SIM,³ the parameters λ , T , K , K_{\max} are set according to the experiment values as introduced in Ref. 3.

The parameters are set according to the existing literatures, the momentum is set to 0.9, and the batch size is set to 50. The learning rate for our model is set to 0.00002 at the start and we drop the learning rate by a factor of 10.

4.1.4. Metrics

In this paper, the performance of different approaches is evaluated by five popular metrics, namely average precision, one-error, hamming loss, ranking loss and coverage. For average precision, the bigger the value the better the performance, it evaluates the average fraction of ranked above a particular label; while for the other four metrics, the smaller the value, the better the performance. More details can be found in Ref. 34. The mentioned five metrics measure the performance from different aspects; it is difficult for one approach to outperform another on every one of these metrics. We first conduct pre-training of CNNs on the training datasets to obtain the parameters.

4.2. Results

4.2.1. Evaluation on MIRFLICKR-25000 dataset

As discussed before, the images in this data are weakly labeled. On this data, we only conduct experiments in the image-level model. We compare the performance of MIMLCNNs with MIMLBOOST,³⁴ MIMLSVM,³⁴ SISL-MIML,¹⁷ M3MIML³² and

Table 4. The performance of different approaches on the MIRFLICKR-25000 dataset.

Approach	Average Precision	One Error	Hamming	Ranking	Coverage
MIMLCNNs	0.618	0.264	0.199	0.257	1.261
MIMLBOOST	0.473	0.328	0.263	0.281	1.297
MIMLSVM	0.502	0.307	0.277	0.303	1.567
MIMLKNN	0.481	0.311	0.251	0.288	1.422
M ³ MIML	0.529	0.289	0.247	0.279	1.439
SISL-MIML	0.588	0.363	0.232	0.1708	1.198

MIMLKNN.³⁰ The performance of each compared approach is evaluated by conducting five-fold cross-validation on the MIRFLICKR-25000 dataset.

In our network, an image is first sampled from the training dataset before feeding the images in triplets to the CNNs, each image is resized and a 224*224 patch at random position from each image is cropped as input. This provides an augmentation of the dataset which is demonstrated to improve the generalization of the network.

Table 4 shows the results of our deep MIML learning approach on MIRFLICKR-25000 dataset. We measure the performance in the above-mentioned five metrics. We can see from Table 4 that the proposed approach yields highly encouraging performance in terms of average precision, one error, ranking loss and Hamming loss while it achieves superior performance to other approaches on the other evaluation metrics. By processing the images directly, MIMLCNNs outperform the compared approaches which treat semantic labels separately and neglect the interactions among them. Especially in one error, MIMLCNN shows a particularly prominent advantage, compared with the traditional optimal approach M³MIML, our approach improves the result by 8.65%. In summary, the performance results show MIMLCNN is an effective approach to the task of MIML problem.

Moreover, we also investigate the behavior of different approaches as we vary the number of training examples. We select a certain proportion of data as the training set and the remaining data as the testing set. Figure 4 shows the performance of MIMLCNNs and other four approaches MIMLBOOST,³⁴ MIMLSVM,³⁴ M³MIML³² and MIMLKNN.³⁰ Although MIMLBOOST³⁴ and M³MIML³² are time-consuming, they are included. In order to compare conveniently, we plot the results of 1-average precision instead of average precision, the lower the curve, the better is its performance. We can see from Fig. 4 that MIMLCNNs outperforms MIMLBOOST,³⁴ MIMLSVM³⁴ and M³MIML³² in general, especially when large percentage ratio of the dataset is employed.

Some of the results are demonstrated in Table 5. For each of these images, the main three tags calculated based on the proposed approach are provided.

In Table 5, we can observe that the main results are in agreement with the raw labels, especially the first label. So, the value of one error metric is very effective, the reason is that users usually focus on a few main instances in an image, especially the prominent one; the main label can well represent the image. As a result, the

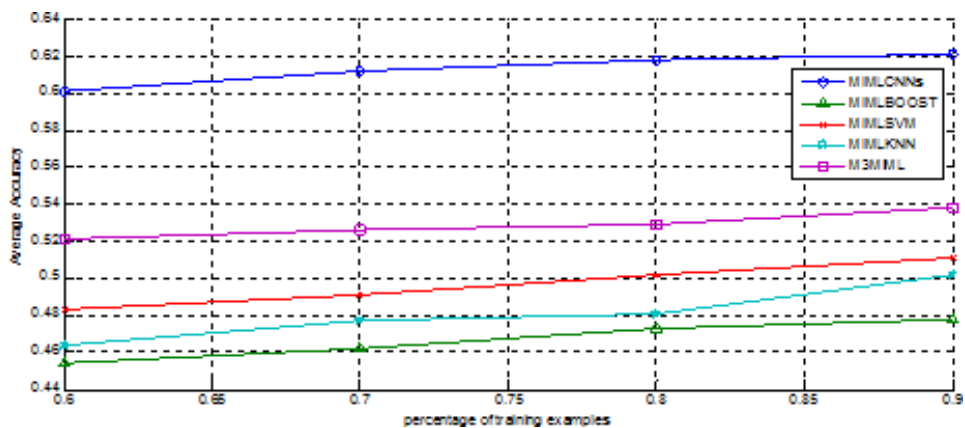


Fig. 4. The performance of different approaches with respect to the number of training examples.

probability is relatively small that the highest confidence in the prediction of the label does not belong to the images.

4.2.2. Evaluation on MSRC dataset

Here, we show the application of our framework in image annotation. In this section, the proposed approach is conducted on the dataset to perform both image-level and instance-level annotations. MIMLCNNs is compared with MIMLBOOST,³⁴ MIMLSVM,³⁴ MIMLKNN,³⁰ and SIM.³ SIM³ is a MIML approach which minimizes ranking loss for instance-level prediction. The performance of each approach is evaluated by conducting five-fold cross-validation on the MSRC dataset. We randomly select 80% of images as the training set with an additional constraint that it should contain at least half positive images of each class. The remaining 20% is the testing set.

Table 5. Example outputs produced by the MIMLCNNs approach.



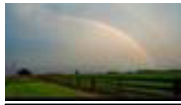
Image	Raw labels	The MIMLCNNs results
	Rainbow; Tropical;Avcation; Wil wheatoo	Rainbow; clouds; tree
	Rainbow;Fjord;Nissu;Path; Explore;pot of gold; bifront Zoog;regard	Rainbow;sky;clouds
	NIKON;D40x;10-20 mm; Columbum;Clouds; Rainbow Dublin	Rainbow;clouds;sky

Table 6. The performance of different approaches on the MSRC dataset.

Approach	Average Precision	One Error	Hamming	Ranking	Coverage
MIMLCNNs	0.782	0.292	0.083	0.098	0.209
MIMLBOOST	0.684	0.328	0.107	0.119	0.255
MIMLSVM	0.685	0.334	0.084	0.121	0.259
MIMLKNN	0.599	0.437	0.131	0.158	0.319
SISL-MIML	0.687	0.316	0.110	0.107	0.243

Table 6 shows the experiment results of the five compared approaches. For each evaluation criterion, it is obvious that MIMLCNN performs better than the other four existing approaches. MIMLCNN achieves average precision and coverage improvement of around 30.55% and 34.48% compared with MIMLKNN.³⁰ Specifically, the average precision and coverage are far superior to the compared approaches due to the deep representations.

In addition, we investigate the behavior of different approaches as we vary the size of training examples. We randomly pick up 60%, 70%, 80% and 90% of the MSRC dataset as the training set, the remaining data as the testing set. We can see from Fig. 5 that MIMLCNN achieves the best performance in most cases.

In addition, because pixel-level labels are included, the MSRC dataset can be useful for the instance-level annotation problem. An image example and its' pixel are shown in Fig. 6. The image is divided into five instances: road, sky, tree, building and car.

In this paper, we conduct experiment on instance-level by five-fold cross-validation. Through learning the training set to predict the unknown labels. MIMLCNN is compared with M3MIML³² and SIM.³ M3MIML is image-level approach, but it uses the instance-level model to train, so it can be used to predict labels. The parameters are set according to the values mentioned in Sec. 4.1.3. Table 7 shows the experiment results of the compared approaches. SIM³ calculates the maximum prediction value

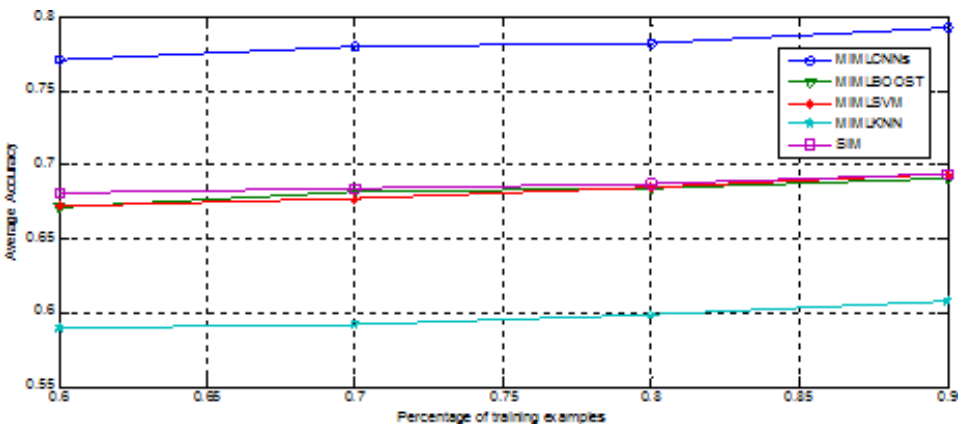


Fig. 5. The performance of different approaches with respect to the number of training examples.



Fig. 6. An example and corresponding instances.

Table 7. Results of the compared approaches.

Approach	Average Precision
MIMLCNNs	0.712
M ³ MIML	0.547
SIM	0.644

to determine the corresponding relationship between images and labels. However, we can observe from the table that MIMLCNN achieves convincing performance than SIM.³ The average precision improves about 7.23%.

5. Conclusion

In this paper, we proposed to construct a deep learning framework within a supervised learning setting. We demonstrate that the deep MIML learning approach that we developed performs well in image annotations. The proposed approach is also able to automatically extract correspondences between instance and labels and return meaningful label pairs. We hope the findings can improve the research of deep learning and weakly supervised learning. Experiments on real-world datasets are able to compare the proposed approach with existing approaches. Experimental results show the MIMLCNN approach achieves better performance than the existing approaches. In the future work, we want to use other deep learning methods which may obtain competitive results from this model in the MIML learning setting compared to our approach.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 61375121, 61432008, and 61075049), the Provincial Projects of Natural Scientific Research Fund for Jiangsu Universities (No. 14KJD520003), and

sponsored by the Scientific Research Funds of Jinling Institute of Technology for Introducing Talents (No. jit-rcyj-201505).

References

1. K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei and M. I. Jordan, Matching words and pictures, *JMLR* **3** (2003) 1107–1135.
2. M. R. Boutell, J. Luo, X. Shen and C. Brown, Learning multi-label scene classification, *Pattern Recogn.* **37**(9) (2004) 1757–1771.
3. F. Briggs, X. Z. Fern, R. Raich and Q. Lou, Instance annotation for multi-instance multi-label learning, *ACM Transactions on Knowledge Discovery from Data* **7**(3) (2013) 1–30.
4. G. Carneiro, A. B. Chan, P. J. Moreno and N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, **29**(3) (2007) 394–410.
5. A. Fakeri-Tabrizi, M. R. Amini and P. Gallinari, Multiview semisupervised ranking for automatic image annotation, in *Proc. ACM Int. Conf. Multimedia (ACMMM)* (ACM, 2013), pp. 513–516.
6. R. Fergus, Y. Weiss and A. Torralba, Semi-supervised learning in gigantic image collections, in *Advances in Neural Information Processing Systems (NIPS)* (2009) 522–530.
7. S. Gao, Z. Wang, L.-T. Chia and I. W.-H. Tsang, Automatic image tagging via category label and web data, in *Proc. ICM* (2016) 1115–1118.
8. L.-H. Guo and L.-W. Jin, The generic object classification based on MIML machine learning, in *Chinese Conference on Pattern Recognition (CCPR2009)* (Nanjing, November 4–6, 2009).
9. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, *CoRR*, Vol. abs/1207.0580, (2012). Available at <http://arxiv.org/abs/1207.0580>.
10. A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems (NIPS)* (2012), pp. 1097–1105.
11. Y.-X. Li, Drosophila gene expression pattern annotation through multi-instance multi-label learning, *IEEE Trans. Comput. Biol. Bioinfo.* **9**(1) (2015) 98–112.
12. J. Li and J. Z. Wang, Real-time computerized annotation of pictures, *IEEE TPAMI* **30**(6) (2008) 985–1002.
13. Y. Li, Z. Qi, Z. M. Zhang and M. Yang, Learning with limited and noisy tagging, in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2015, pp. 957–966. Available at <http://doi.acm.org/10.1145/2502081.2502111>.
14. M. Lin, Q. Chen and S. Yan, Network in network, in *Int. Conf. Learning Representations (ICLR)* (2013), pp. 1–10.
15. Z. Ma, Y. Yang, F. Nie, J. Uijlings and N. Sebe, Exploiting the entire feature space with sparsity for automatic image annotation, in *Proc. ACM Int. Conf. on Multimedia (ACMMM)* (ACM, 2011), pp. 283–292.
16. F. Monay and D. Gatica-Perez, Plsa-based image auto-annotation: Constraining the latent space, in *Proc. ACM Int. Conf. Multimedia (ACMMM)* (ACM, 2014), pp. 348–351.
17. N. Nguyen, A New SVM Approach to Multi-instance multi-label learning, in *IEEE 10th International Conference on Data Mining (ICDM2010)* (Sydney, NSW, December 13–17, 2010).

18. H. O. Song, Y. J. Lee, S. Jegelka and T. Darrell, Weakly supervised discovery of visual pattern configurations, in *NIPS*, 2014.
19. C. Szegedy *et al.*, Going deeper with convolutions, CoRR, Vol. abs/1409.4842, 2014. [Online]. Available: <http://arxiv.org/abs/1409.4842>.
20. X. Tan, F. Wu, X. Li, S. Tang, W. Lu and Y. Zhuang, Structured visual feature learning for classification via supervised probabilistic tensor factorization, *Multimedia, IEEE Trans.* **17**(5) (2015) 660–673.
21. Y. Ushiku, T. Harada and Y. Kuniyoshi, Efficient image annotation for automatic sentence generation, in *Proc. ACM Int. Conf. Multimedia (ACMMM)* (ACM, 2015), pp. 549–558.
22. L. Wan, M. Zeiler, S. Zhang, Y. L. Cun and R. Fergus, Regularization of neural networks using dropconnect, in *Proc. 30th Int. Conf. Machine Learning (ICML)* (2013), pp. 1058–1066.
23. G. Wang and D. Forsyth, Joint learning of visual attributes, object classes and visual saliency, in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Vol. 30 (2009), pp. 537–544.
24. F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang and Y. Zhuang, Sparse multi-modal hashing, *Multimedia, IEEE Trans.* **16**(2) (2014) 427–439.
25. X. Xu and E. Frank, Logistic regression and boosting for labeled bags of instances, in *Lecture in Artificial Intelligence*, eds. H. Dai, R. Srikant and C. Zhang (Springer, Berlin, 2004) pp. 272–281.
26. X.-S. Xu, Y. Jiang, L. Peng, X. Xue and Z.-H. Zhou, Ensemble approach based on conditional random field for multi-label image and video annotation, in *Proc. ACM Int. Conf. Multimedia (ACMMM)* (ACM, 2015), pp. 1377–1380.
27. Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, E. I. Chang *et al.*, Deep learning of feature representation with multiple instance learning for medical image analysis, in *ICASSP*, 2014.
28. X.-Y. Xue, W. Zhang and B. Wu, Correlative multi-label multi-instance image annotation, in *IEEE Conf. Computer Vision (ICCV2011)*, (Barcelona, November 6–13, 2011).
29. J.-J. Yan and Y.-Q. Wang, A multi-instance multi-label learning approach to objective auscultation analysis of traditional Chinese medicine, in *The 4th Int. Conf. Biomedical Engineering and Informatics (BMEI2011)*, (Shanghai, October 15–17, 2011).
30. M.-L. Zhang, A k -nearest neighbor based multi-instance multi-label learning algorithm, in *IEEE Conf. Tools with Artificial Intelligence* (Arras, France, 2016), pp. 207–212.
31. M.-L. Zhang and Z.-H. Zhou, A k -nearest neighbor based algorithm for multi-label classification, *IEEE Conf. Granular Computing*, Beijing, China, July 25–27, 2005.
32. M.-L. Zhang and Z.-H. Zhou, M3-MIML: A maximum margin method for multi-instance multi-label learning, in *IEEE 8th International Conference on Data Mining (ICDM2008)* (Pisa, December 15–19, 2008).
33. Z.-H. Zhou, Multi-instance learning: A survey, Technical Report, Nanjing University, 2004.
34. Z.-H. Zhou and M.-L. Zhang, Multi-Instance multi-label learning with application to scene classification, in *Advances in Neural Information Processing Systems 19, Proc. 20th Annual Conf. Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 4–7 (2006), pp. 1609–1616.



Hai-Feng Guo received her M.S. degree in Pattern Recognition in 2009 from the University of Hohai in Nanjing, China. She is currently a Ph.D. student at the Computer Science Department of the University of Hohai in Nanjing, China. She was a

lecturer in the School of Computer Engineering, Jinling Institute of Technology in Nanjing, China. Her research interests include image retrieval, pattern recognition as well as recommended system.



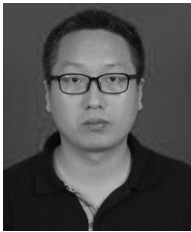
Lixin Han received his Ph.D. in Computer Science from Nanjing University, Nanjing, China. He has been a Post-Doctoral Fellow with the Department of Mathematics, Nanjing University, and a Research Fellow with the Department of Electronic

Engineering, City University of Hong Kong, Kowloon, Hong Kong. He is currently a Professor at the Institute of Intelligence Science and Technology, Hohai University, Nanjing, China. He has published over 60 research papers. Prof. Han is an invited reviewer for several renowned journals and has been a Program Committee Member of many international conferences. He is listed in Marquis' Who's Who in the World and Marquis' Who's Who in Science and Engineering.



Shoubao Su received his Ph.D. and M.S. degrees in Computer Application Technology both from Anhui University, China. Currently, he is a Professor at the School of Computer Engineering at Jinling Institute of Technology, China. His research

interests focus on applied swarm intelligent computing to big data mining, pattern recognition and embedded control optimization.



Zhou-Bao Sun received his B.S. degree in Math and Computer Science in 2007 from the Math Department of Anhui University of Science and Technology in Anhui, China. He received his Ph.D. in Computer Application Technology in

2015 from the University of Hohai in Nanjing, China. He research interests include pattern recognition, image processing, machine learning and data mining.